This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2896779, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

*Digital Object Identifier 10.1109/ACCESS.2017.DOI*

# Video-based Person Re-Identification Using Unsupervised Tracklet Matching

**CHIRINE RIACHY[1], FOUAD KHELIFI[1], (Member, IEEE), AHMED BOURIDANE[1], (Senior Member, IEEE)**

[1]Department of Computer and Information Sciences, Northumbria Univeristy, Newcastle upon Tyne, UK

Corresponding author: Ahmed Bouridane (e-mail: ahmed.bouridane@northumbria.ac.uk).

This publication was made possible by NPRP grant No. 8-140-2-065 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

**ABSTRACT** Despite the significant improvement in accuracy supervised learning has brought into person re-identification (re-id), the availability of sufficient fully annotated data from concerned camera-views poses a problem for real-life applications. To alleviate the burden of intensive data annotation, one way is to resort to unsupervised methods. This has motivated us to propose a novel algorithm for unsupervised video-based person re-id applications. To achieve this, the frames of a person video tracklet are divided into a set of clusters that are subsequently matched using a distance measure based on the Naive Bayes Nearest Neighbor algorithm and Spearman distance. Knowing that person sequences may suffer from substantial changes in viewpoint, pose and illumination distortions, our technique allows the rejection of poor and noisy clusters while retaining the most discriminative ones for matching. Experiments on three widely used datasets for video person re-id PRID2011, iLIDS-VID and MARS have been carried out, and the results demonstrate the superiority of the proposed approach.

**INDEX TERMS** Person re-identification, Spearman distance, unsupervised method, video surveillance.

## I. INTRODUCTION

Matching people across cameras is of great interest especially for security applications. When a query person is presented, retrieving that person from a gallery of people captured under a different camera view is known as person re-identification (re-id). In the case where subjects are represented by video sequences, the problem of video-based person re-id is encountered.

The past few years have witnessed a large focus on metric learning [1]–[4] and deep learning [5]–[14] to solve the re-id problem. These methods have largely contributed into the advancement of the field by considerably boosting the performance. However, most of these methods require the availability of a sufficient amount of annotated data from concerned camera views to train the model before re-id can take place. This still is a hindrance for the applicability of re-id systems into real-world problems. In addition to the high annotation cost, the availability of enough matched instances under the camera views in question is a requirement that is not easily fulfilled. These reasons motivate our work towards improving unsupervised video person re-id to move a step closer into solving the real-life problem.

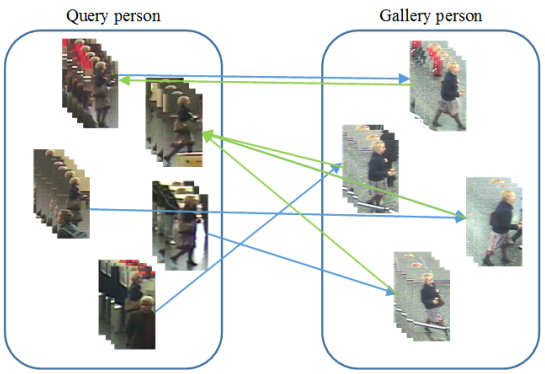


**FIGURE 1.** A set-based matching process allows the selection of better representative frames to be associated with their counterparts in the gallery set.

A common practice to address video-based or multi-shot re-id is by adapting the task into the single-shot scenario. This could be achieved by averaging or max-pooling frame-wise feature vectors of a person sequence to obtain one final vector for each person tracklet [15]–[18]. In this case, a single-shot matching method can subsequently be adopted

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2896779, IEEE Access



to measure the similarity between probe and gallery vectors. Although such approach is simple and efficient, it suffers from major drawbacks. Firstly, the final representation of a person sequence is biased towards the pose that is most common among the frames constituting that sequence. Due to disjoint camera views causing significant illumination, pose and viewpoint angle variations [16], [19], this will frequently result in large intra-class variations between positive matches which will degrade the performance. Secondly, such an approach treats noisy outliers with the same importance as the good informative frames by assigning them all equal weights.

To address the disadvantages presented by feature pooling, we propose a simple yet robust learning-free distance measure based on the Naive Bayes Nearest Neighbor (NBNN) classifier [20]. The latter has been initially used in image classification in the context of Image-to-Class distance. However, in this work, we leverage NBNN algorithm to formulate a Set-to-Set distance measure. We also integrate Spearman rank correlation coefficient into this framework as the similarity kernel achieving remarkable improvement. Nearest neighbor based classifiers have been successfully used in image classification [20], [21], action recognition [22], [23] and most recently they also have been investigated in person re-id problems [24], [25]. However, to the best of our knowledge, NBNN based methods and correlation type distances such as Spearman distance have not been investigated in re-id yet.

Our contributions are as follows: (i) We extend a state-of-the-art image descriptor into 3 dimensions achieving significant improvement in accuracy. (ii) By regarding person re-id as a classification problem, we formulate the multi-shot re-id task as a set-based matching problem that we tackle using the NBNN classifier. (iii) For the first time, we explore Spearman correlation distance based on rank vectors as opposed to common distance metrics such as the Euclidean distance. (iv) Finally, we advance the state-of-the-art significantly for unsupervised re-id on two challenging datasets PRID2011 and iLIDS-VID. We also achieve competitive results on MARS dataset. The improvement is over 22% and 6% in rank-1 accuracy for iLIDS-VID and PRID2011 respectively, compared to the current best performing method [15].

## II. RELATED WORK

**Video-based person re-id.** The recent popularity of video-based person re-id [8], [12], [26]–[32] is motivated by two main reasons. Firstly, person videos are generally available from surveillance cameras which makes the video re-id problem a more realistic one. Secondly, video sequences enable the use of temporal cues and provide rich person representations due to the availability of multiple images for each person. A few methods have therefore been proposed. Early trends focused on designing spatio-temporal descriptors [33], [34] before matching methods [27], [31], [33], [35] were explored. In more recent years, especially with the release of large-scale datasets [18], deep learning gained huge popularity among re-id researchers [8], [12]–[14], [29], [30], [32] who proposed different deep architectures to tackle the challenges associated with cross-view matching in person re-id. Nonetheless, a major drawback persists with these algorithms that renders their use in real-life scenarios almost inapplicable. Specifically, most of these methods require large-scale annotated datasets from pairwise camera-views for model training. This is not only prohibitively costly, but this information is most likely unavailable given a specific pair of cameras. This encourages resorting to unsupervised methods to circumvent this requirement.

**Unsupervised video re-id.** Although unsupervised person re-id is not extensively researched compared to the supervised problem, the attention of the re-id community seems to be shifting towards this direction lately [6], [15], [24], [36]–[38]. Solving the unsupervised video re-id task is a massive step towards real-world deployment. For this purpose, a few algorithms have been recently proposed. For instance, Liu *et al.* [34] proposed a spatio-temporal person descriptor by extracting walking cycles and body action units that are subsequently described using Fisher vectors. Ma *et al.* [39] developed a video representation based on spatio-temporal pyramids and performed sequence matching using a modified dynamic time warping algorithm. Liu *et al.* [24] and Ye *et al.* [37] devised methods to estimate the labels progressively through iterative algorithms with hand-crafted features. However, they implicitly [24] or explicitly [37] used labels from one camera view for model initialization. More recently, Chen *et al.* [15] and Li *et al.* [6] explored end-to-end deep learning architectures to associate within-camera and cross-camera tracklets by optimizing specifically tailored objective functions.

In this work, we propose a fully unsupervised video re-id method using a robust spatio-temporal descriptor. We also design a set-based matching method by leveraging the NBNN algorithm, to which we incorporate a correlation type distance based on rank vectors, Spearman distance. Despite its simplicity, the proposed system closes the gap with supervised methods on two widely used benchmarks and produces very competitive results on a large-scale public dataset.

## III. PROPOSED APPROACH

Considering the large amount of video data available from surveillance cameras, existing pedestrian detection and tracking algorithms [40], [41] can be readily employed to extract person tracklets. For the supervised video re-id task, person tracklets are collected from disjoint cameras, and each person is assigned an ID. Accordingly, within- and between-camera tracklets are annotated. Although in this work each person tracklet is treated separately for feature extraction and clustering as will be explained in the sequel, no assumption on the identity of the person involved is made. Therefore, unlike previous works that require labels from one camera view for model initialization [24], [37], the proposed technique is purely unsupervised, and does not involve any learning. A brief description is given as follows. Firstly, high-dimensional GOG3D features are extracted from person

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2896779, IEEE Access



tracklets. The dimension of these features is then reduced using PCA algorithm. The obtained frame-wise features of each tracklet are subsequently clustered using k-means algorithm, and each cluster is represented by its centroid. Finally, an NBNN-based distance measure is used to compute probe-gallery distances.

At the moment of testing, when a probe person tracklet is presented to the system, pairwise distances between probe and all gallery tracklets are computed. Gallery elements are ranked according to their distance from the probe. The correct match(es) should ideally appear in high rank(s). The components of the proposed approach including person representation and matching method are detailed thereafter. A representative diagram is shown in Fig. 2.

### A. FEATURES

For person sequences representation, we extend the state-of-the-art GOG descriptor [42] into 3 dimensions, we denote it GOG3D. The main modification lies in the pixel feature vector that is now represented by $f = [x, y, M_{0^\circ}, M_{90^\circ}, M_{180^\circ}, M_{270^\circ}, |I_t|, L, A, B]^T$ where $x$ and $y$ as the x- and y-coordinates of the pixel, $M_{0^\circ}$ through $M_{270^\circ}$ are the bins into which the gradient orientation is quantized and multiplied by the gradient magnitude, $|I_t|$ is the magnitude of the temporal gradient, and $L$, $A$ and $B$ are the LAB color channels. Similarly to GOG, images are first divided into small overlapping patches and $R$ horizontal regions. Each patch is initially modeled by a Gaussian, and patches in the same horizontal region are subsequently summarized by a unique Gaussian to achieve some viewpoint invariance. These region Gaussians are then projected into the Euclidean space and concatenated to form the final image representation. More details on these steps can be seen in [42].

Mean removal and $\ell_2$-normalization are finally applied to the extracted features, and the dimension is reduced by Principal Components Analysis (PCA). Employing GOG3D instead of GOG contributes mainly to leveraging temporal information in addition to color and texture cues for better discriminability between persons.

Matching between frame-wise features directly is not only costly, it also ignores the rich representation obtained by combining information from various frames. For this purpose, we cluster the PCA-reduced frame-wise features constituting each video tracklet using k-means algorithm, and we represent each cluster by its centroid. That yields a set of feature vectors for each video tracklet rendering the probe-gallery matching a set-based process.

### B. NBNN-BASED DISTANCE MEASURE

Let $P = (p_1, \ldots, p_n)$ be a probe person sequence represented using $n$ feature vectors (clusters centroids) obtained by clustering frame-wise features. Person re-id aims at finding the class in the gallery set to which this probe belongs. Namely, the purpose is to find the gallery subject $G_k = (g_1^k, \ldots, g_m^k)$ that is a correct match for query person $P$.

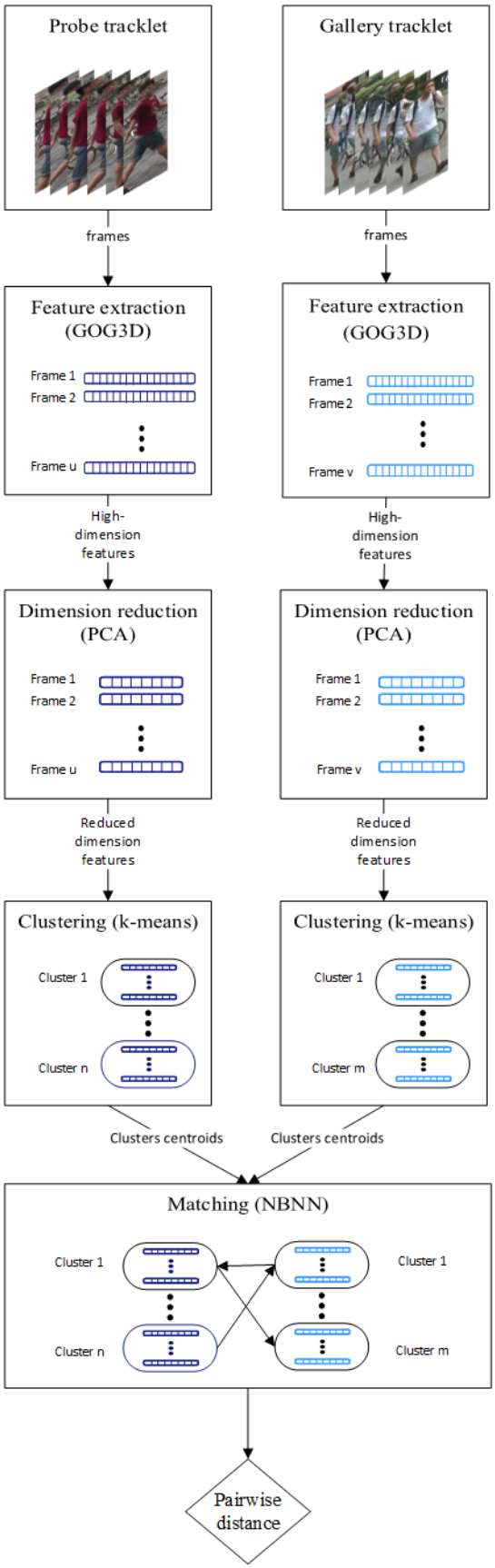


**FIGURE 2.** Diagram of the proposed method.

The maximum-a-posteriori (MAP) classifier minimizes the average classification error:

$$\hat{C} = \underset{k}{\operatorname{argmax}}\, p(G_k|P), \tag{1}$$

where $k \in \{1, \ldots, C\}$, $C$ being the number of classes in the gallery set. The following Bayes' rule can subsequently be applied:

$$p(G_k|P) = \frac{p(G_k).p(P|G_k)}{p(P)}. \tag{2}$$

Assuming a uniform prior over classes $G_k, k \in 1, \ldots, C$, and $p(P)$ being a constant independent of the class $G_k$, MAP classifier reduces into the maximum-likelihood (ML) classifier as such:

$$\hat{C} = \underset{k}{\operatorname{argmax}}\, p(G_k|P) = \underset{k}{\operatorname{argmax}}\, p(P|G_k). \tag{3}$$

Assuming that probe descriptors $p_1, \ldots, p_n$ satisfy the Naive Bayes assumption (they are i.i.d. given class $G_k$), $p(P|G_k)$ can be written as:

$$p(P|G_k) = p(p_1, \ldots, p_n|G_k) = \prod_{i=1}^{n} p(p_i|G_k). \tag{4}$$

Substituting $p(P|G_k)$ from (4) in (3) and taking the log probability yields:

$$\hat{C} = \underset{k}{\operatorname{argmax}} \log p(P|G_k) = \underset{k}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(p_i|G_k). \tag{5}$$

If $g_1^k, \ldots, g_m^k$ are all the descriptors in class $G_k$, then $p(p_i|G_k)$ can be approximated using a Parzen window estimator [21] with similarity kernel $K$ by:

$$\hat{p}(p_i|G_k) = \frac{1}{m} \sum_{j=1}^{m} K(p_i - g_j^k), \tag{6}$$

$\hat{p}(p_i|G_k)$ in (6) can be further approximated by taking the $r$ largest elements in this summation. They correspond to the $r$ nearest neighbors of $p_i$ in class $G_k$:

$$\hat{p_r}(p_i|G_k) = \frac{1}{m} \sum_{j=1}^{r} K(p_i - g_j^k). \tag{7}$$

This can be taken to the extreme by using a single nearest neighbor of $p_i$ in $G_k = \{g_1^k, \ldots, g_m^k\}$ denoted $NN_{G_k}(p_i)$. Hence,

$$\hat{p_1}(p_i|G_k) = \frac{1}{m} K(p_i - NN_{G_k}(p_i)). \tag{8}$$

Choosing a single nearest neighbor is particularly appealing because in that case, equation (5) can reduce into a very simple format [21]. For instance, by selecting a Gaussian kernel for $K$ and combining equations (5) and (8) we obtain:

$$\begin{aligned} \hat{C} &= \underset{k}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(p_i|G_k) \\ &= \underset{k}{\operatorname{argmax}} \sum_{i=1}^{n} \log(\frac{1}{m} K(p_i - NN_{G_k}(p_i)) \\ &= \underset{k}{\operatorname{argmax}} \sum_{i=1}^{n} \log(\frac{1}{m} e^{-\frac{1}{2\sigma^2}||p_i - NN_{G_k}(p_i)||^2}) \\ &= \underset{k}{\operatorname{argmin}} \sum_{i=1}^{n} ||p_i - NN_{G_k}(p_i)||^2 \end{aligned} \tag{9}$$

In other terms, the NBNN classification rule entails finding the gallery instance with the minimum distance from the probe. Based on the previous analysis, noting that in our case $n = m = 5$ (number of k-means clusters) for all probe and gallery elements, we define the distance between a probe instance $P = (p_1, \ldots, p_n)$ and a gallery instance $G = (g_1, \ldots, g_m)$ as:

$$d_{P \to G} = \sum_{i=1}^{n} \delta(p_i, NN_G(p_i)), \tag{10}$$

where $NN_G(p_i)$ is the nearest neighbor of $p_i$ in $G$, and $\delta$ is a distance measure such as the Euclidean or Cosine distance. Similarly, the distance from $G$ to $P$ can be computed as:

$$d_{G \to P} = \sum_{i=1}^{m} \delta(g_i, NN_P(g_i)), \tag{11}$$

and the final similarity score between $P$ and $G$ combines both formulas in a symmetric manner,

$$d_{NN}(P, G) = d_{P \to G} + d_{G \to P}. \tag{12}$$

Based on this definition, the correct match for a probe $P$ consists of the gallery element $G$ with the minimum distance from $P$. It is worth noting here that a more general derivation incorporating more than one nearest neighbor for each descriptor $p_i$ with arbitrary values for $n$ and $m$ is given by:

$$d_{P \to G} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{r} \delta(p_i, g_j), \tag{13}$$

where $g_j, j = 1, \ldots, r$ are the $r$ nearest neighbors of $p_i$ in class $G$. Similarly,

$$d_{G \to P} = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{r} \delta(g_i, p_j), \tag{14}$$

where $p_j, j = 1, \ldots, r$ are the $r$ nearest neighbors of $g_i$ in probe $P$. Eventually, the final distance $d_{NN}$ is derived as in (12).

### C. EMBEDDED MATCHING DISTANCE

The embedded distance $\delta$ plays a crucial role in the success of the proposed re-id system. Although the Euclidean distance has been widely used with unsupervised person re-id methods [15], [43], we believe it might not be the best option.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2896779, IEEE Access



By treating all features equally, the Euclidean distance is very vulnerable to outliers. For instance, notable fluctuations in a few features might affect the final pairwise distance drastically. As these are likely to happen in the re-id scenario due to cross-view camera variations and poor quality images, we propose to use a rank-based distance measure instead. For this purpose, Spearman distance is exploited. It is defined as the Pearson correlation distance applied to rank vectors. That is, feature vectors $X = (x_1, \ldots, x_d)$ and $Y = (y_1, \ldots, y_d)$ are converted into rank vectors $r_X = (r_1^X, \ldots, r_d^X)$ and $r_Y = (r_1^Y, \ldots, r_d^Y)$ by replacing all the values by their respective ranks. For instance, if vector $X = (0.7, 0.2, 0.4)$, then $r_X = (3, 1, 2)$. Subsequently, Spearman distance between $X$ and $Y$ is computed in terms of Spearman rank correlation coefficient $\rho_s$ as follows:

$$\begin{aligned} d_S(X,Y) &= 1 - \rho_s(X,Y) \\ &= 1 - \frac{\sum_{i=1}^{d}(r_i^X - \overline{r_X})(r_i^Y - \overline{r_Y})}{\sqrt{\sum_{i=1}^{d}(r_i^X - \overline{r_X})^2}\sqrt{\sum_{i=1}^{d}(r_i^Y - \overline{r_Y})^2}} \end{aligned} \tag{15}$$

where $\overline{r_x} = \frac{1}{d}\sum_{i=1}^{d} r_i^X = \frac{d+1}{2}$ and $\overline{r_Y} = \frac{1}{d}\sum_{i=1}^{d} r_i^Y = \frac{d+1}{2}$ are the means of rank vectors $r_X$ and $r_Y$, respectively.

To further justify the use of Spearman distance, let us start by defining other common distance measures: Euclidean, Cosine, and Pearson correlation distance. Element-wise notation is used for more clarity. If $\overline{x}$ and $\overline{y}$ are the respective means of vectors $X$ and $Y$, then the distances in the previous order are defined as as follows:

$$d_{Euc}(X,Y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2} \tag{16}$$

$$d_{Cos}(X,Y) = 1 - \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2}\sqrt{\sum_{i=1}^{d} y_i^2}} \tag{17}$$

$$d_P(X,Y) = 1 - \frac{\sum_{i=1}^{d}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{d}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{d}(y_i - \overline{y})^2}} \tag{18}$$

Since our features are zero-centered and $\ell_2$-normalized, it can be easily seen that these 3 distance measures become equivalent. Therefore, the comparison of Spearman distance to one of them applies to all. In fact, from the formulations we can see that Spearman distance is Pearson distance applied to rank vectors. By taking ranks instead of raw values, Spearman distance relaxes the assumption held by the former of a linear relationship between feature vectors for higher similarity, and looks for patterns of monotonicity instead. That is, in this case 2 vectors exhibit higher similarity when their features undergo the same type of fluctuations. This proves to be more useful in the case of challenging person re-id. The superiority of Spearman distance compared to other metrics will be shown experimentally in section IV-D.

## IV. EXPERIMENTS AND RESULTS

### A. DATASETS

**PRID2011** dataset [44] contains image sequences captured by two adjacent static surveillance cameras. It includes 385 persons in camera A and 749 in camera B among which 200 subjects appear in both camera views. Sequences lengths range from 5 to 675 frames with an average number of 100 frames per sequence. The challenges associated with this dataset include mainly significant illumination changes and viewpoint angle variations. We follow the common evaluation protocol for PRID2011 [6], [15], [24], [33], [37], [45] by retaining 178 video pairs with more than 27 frames per sequence. Although training is not needed in our case, the dataset is randomly divided into half for training and half for testing (89 persons in each subset) for fair comparison with other methods, and experiments are repeated over 10 trials. Average results are reported using Cumulative Matching Characteristic (CMC) and top-matching rates.

**iLIDS-VID** dataset [33] consists of 600 image sequences for 300 individuals captured by two non-overlapping camera views in an airport arrival hall. Sequences lengths vary from 23 to 192 frames with an average number of 73 frames per sequence. iLIDS-VID is very challenging due to significant cross-view illumination and viewpoint angle variations, occlusions, and background clutter. Similar to PRID2011, iLIDS-VID is randomly partitioned into two equal subsets of 150 persons each for training and testing. Experiments are repeated 10 times and average results in CMC top-matching rates are reported.

**MARS** [18] is a large-scale video benchmark for person re-id collected on a university campus by 6 near-synchronized cameras. It consists of 1,261 pedestrians each appearing in 2 cameras at least. Unlike other benchmarks, persons are not manually detected and cropped. Alternatively, a more realistic application is offered by automatic pedestrian detection and tracking using DPM [40] detector and GMMCP [41] tracker. This results in a total of 20,478 person tracklets with an average of 13.2 tracklets per person, including 3,248 distractors caused by false detection or tracking. The standard train/test split [18] used by other algorithms [6], [15], [24], [37] is adopted for evaluation. One tracklet is selected for each person from each view as probe. As multiple ground truths correspond to each query, in addition to CMC curve, mean Average Precision (mAP) is used to evaluate the performance.

### B. IMPLEMENTATION DETAILS

For the proposed GOG3D feature, we use a different parameter setting from GOG [42]. Particularly, we set the patch size to $9 \times 9$ pixels where patches are extracted at 2 pixels intervals, and we divide the image into 10 horizontal regions with 50% overlap. The regularization parameter to ensure non-singular covariance matrices is set to $\epsilon = 0.0001$. PRID2011 and iLIDS-VID frames are kept in their original size of $128 \times 64$ pixels, and MARS frames are resized to $128 \times 48$ pixels before extracting features for efficiency. This

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2896779, IEEE Access



results in frame-wise feature vectors of 22,780 dimensions each. The number of k-means clusters is set to 5 clusters ($m = n = 5$ in (13) and (14)) for each probe or gallery tracklet in all our experiments. In fact, small difference was empirically observed when varying the number of clusters $k$, therefore a small $k$ is conveniently used for higher computational efficiency. The dimension of the feature is reduced using PCA so that enough components are kept to retain 95% of the variance in the original feature space. One nearest neighbor is used in our experiments (r=1 in (13) and (14)) unless otherwise specified.

### C. COMPARISON TO STATE-OF-THE-ART TECHNIQUES

We compare our method against 9 state-of-the-art unsupervised person re-id methods on PRID2011 and iLIDS-VID datasets, and 5 on MARS dataset. We also report the latest published supervised re-id results to show the current existing gap in performance. STFV3D, MDTS-DTW, unKISS and PAM+LOMO rely on hand-crafted representations with different unsupervised matching methods. SMP and DGM also leverage hand-crafted features, but design algorithms to generate labels that are subsequently used with supervised metric learning. DAL and TAUDL are unsupervised deep models that attempt to associate tracklets in an end-to-end fashion. It is clear from the results reported in Table 2 that the proposed method outperforms all existing unsupervised techniques on PRID2011 and iLIDS-VID datasets by a margin of approximately 6% and 22% respectively in rank-1 accuracy with its closest competitor DAL. It also surpasses supervised methods SDM and QAN, while the gap with the best performing supervised method [26] is almost 6% on iLIDS-VID and less than 2% on PRID2011. The proposed approach also achieves very competitive results on MARS benchmark as can be seen in Table 1. On the latter, it outperforms the 3 non-deep models in rank-1 accuracy and achieves competitive performance with the deep models DAL and TAUDL. However, in general the gap in performance between supervised and unsupervised methods is still very large (approximately 40%) on this dataset.

The poor quality of the bounding boxes produced by automatic detection and tracking on MARS, which causes serious misalignment between consecutive frames as can be seen in Fig. 3, is problematic for hand-crafted features using part-based models (horizontal strips) and temporal cues like GOG3D. Furthermore, it is no surprise that deep learning methods can scale better to large-scale datasets while suffering with small ones. Nonetheless, the amount of data available at the moment of re-identification might not always be substantial, therefore a successful re-id system should be able to strike some balance between both scenarios.

Finally, the proposed approach contributes massively into closing the gap in rank-1 accuracy between supervised and unsupervised re-id on the small but challenging datasets PRID2011 and iLIDS-VID. Meanwhile, more improvement is still required to achieve similar performance for large-scale datasets.

**TABLE 1.** Comparison against state-of-the-art on MARS dataset in top-matching rates and mean Average Precision.

| | Dataset | MARS | | | |
|---|---|---|---|---|---|
| | Rank R | R = 1 | R = 5 | R = 20 | mAP |
| Unsupervised | **DGM+IDE [37]** | 36.8 | 54.0 | 68.5 | 21.3 |
| | **DGM+XQDA [37]** | 23.6 | 38.2 | 54.7 | 11.2 |
| | **SMP [24]** | 23.6 | 35.8 | 44.9 | 10.5 |
| | **DAL (ResNet50) [15]** | **46.8** | **63.9** | **77.5** | 21.4 |
| | **TAUDL [6]** | 43.8 | 59.9 | 72.8 | **29.1** |
| | **Proposed** | 39.7 | 53.2 | 64.1 | 20.1 |
| Supervised | **Snippet [26]** | 86.3 | 94.7 | 98.2 | 76.1 |
| | **STAN [29]** | 82.3 | - | - | 65.8 |
| | **PABR [46]** | 85.1 | 94.2 | 97.4 | 83.9 |
| | **SDM [32]** | 71.2 | 85.7 | 94.3 | - |

### D. ALGORITHM ANALYSIS

In this section, we analyze each component of our system, highlighting the improvement it brings upon the overall performance. For computational reasons, this analysis is conducted solely on PRID2011 and iLIDS-VID.

**GOG vs. GOG3D.** To justify the extension of GOG into 3 dimensions by encoding temporal correlation between consecutive frames, experiments were conducted involving both types of features on PRID2011 and iLIDSVID datasets. The results obtained are shown in Fig. 4. Leveraging temporal information added to the other changes applied to the pixel feature vector and parameter setting by including the x-coordinate and substituting RGB colour channels by LAB channels, contribute into improving the performance by a margin of approximately 3% for PRID2011 and 14% for iLIDS-VID in rank-1 accuracy. The improvement on iLIDS-VID is more remarkable since it additionally suffers from occlusions and significant illumination changes. Hence, leveraging motion information brings additional discriminative information into pedestrian representation.

**PCA vs. no PCA.** In addition to efficiency considerations, reducing the dimension of the feature vectors is essential when no supervision is involved to discard redundant features and select the most discriminative ones for re-identification. This step is particularly important in our case because rank vectors are used for matching. Ranking a high-dimensional feature vector where values slightly differ from each other is sub-optimal. Therefore, to gauge the effect this might have on the system's performance, we evaluate our system with and without PCA. It is noteworthy that the final feature dimension is not user-defined but automatically obtained by retaining 95% of the variance (Section IV-B).

The results obtained are shown in Fig. 5. As expected, employing PCA before matching brings substantial improvement in rank-1 accuracy especially for iLIDS-VID dataset (around 40%). Given the challenges associated with this dataset, selecting discriminative features is essential to match rank feature vectors of the same individual.

**Embedded distance.** To evaluate the effect of the embedded distance on re-id performance, results using 2 distance

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2896779, IEEE Access



**FIGURE 3.** Example tracklets from MARS dataset. Significant misalignment between consecutive frames can be observed which affects the system's performance.

**TABLE 2.** Comparison against state-of-the-art results on PRID2011 and iLIDS-VID datasets in top-matching rates.

| | Dataset | iLIDS-VID | | | PRID2011 | | |
|---|---|---|---|---|---|---|---|
| | Rank R | R = 1 | R = 5 | R = 20 | R = 1 | R = 5 | R = 20 |
| Unsupervised | **STFV3D [34]** | 37.0 | 64.3 | 86.9 | 42.1 | 71.9 | 91.6 |
| | **MDTS-DTW [39]** | 31.5 | 62.1 | 82.4 | 41.7 | 67.1 | 90.1 |
| | **unKISS [45]** | 38.2 | 65.7 | 84.1 | 59.2 | 81.7 | 96.1 |
| | **PAM+LOMO [47]** | 33.3 | 57.8 | 80.5 | 70.6 | 90.2 | 97.1 |
| | **DGM+IDE [37]** | 36.2 | 62.8 | 82.7 | 56.4 | 81.3 | 96.4 |
| | **DGM+XQDA [37]** | 31.3 | 55.3 | 83.4 | 82.4 | 95.4 | 99.8 |
| | **SMP [24]** | 41.7 | 66.3 | 80.7 | 80.9 | 95.6 | 99.4 |
| | **DAL (ResNet50) [15]** | 56.9 | 80.6 | 91.9 | 85.3 | **97.0** | **99.6** |
| | **TAUDL [6]** | 26.7 | 51.3 | 82.0 | 49.4 | 78.7 | 98.9 |
| | **Proposed** | **79.1** | **93.5** | **97.5** | **91.7** | 96.7 | 98.7 |
| Supervised | **Snippet [26]** | 85.4 | 96.7 | 99.5 | 93.0 | 99.3 | 100.0 |
| | **QAN [30]** | 68.0 | 86.8 | 97.4 | 90.3 | 98.2 | 100.0 |
| | **STAN [29]** | 80.2 | - | - | 93.2 | - | - |
| | **SDM [32]** | 60.2 | 847 | 95.2 | 85.2 | 97.1 | 99.6 |

metrics Cityblock ($\ell_1$ distance) and Euclidean are also reported. As previously mentioned, since the extracted features are mean-centered and $\ell_2$-normalized, Cosine distance and Pearson correlation distance are similar to the Euclidean distance (this was indeed verified experimentally). As can be seen in Fig. 6, although the Euclidean distance consistently outperforms Cityblock distance, the improvement in accuracy with Spearman distance is substantial on both datasets. Looking for monotonicity instead of linearity proved in fact useful by considerably improving the system's accuracy.

**Feature pooling vs NBNN.** We also compare our method against a commonly used matching protocol for video person re-id, average-pooling [6], [16], [17]. In that case, frame-wise feature vectors of a tracklet are averaged to obtain one representation of each person sequence. The single-shot scenario with Spearman distance is subsequently used for matching. The results obtained are shown in Fig. 7. Despite that the matching method does not seem to be the most important component of our system, NBNN still brings some consistent improvement over feature-pooling especially in rank-1 accuracy on both datasets.

**Number of nearest neighbors.** It has been argued previously that the number of nearest neighbors considered in the NBNN algorithm barely affects the system's performance [20]. This was also validated experimentally as can be seen in Fig. 8, where the number of nearest neighbors (parameter r) was varied between 1 and 5. Small difference or slightly worse performance is observed for bigger r, which encour-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2896779, IEEE Access



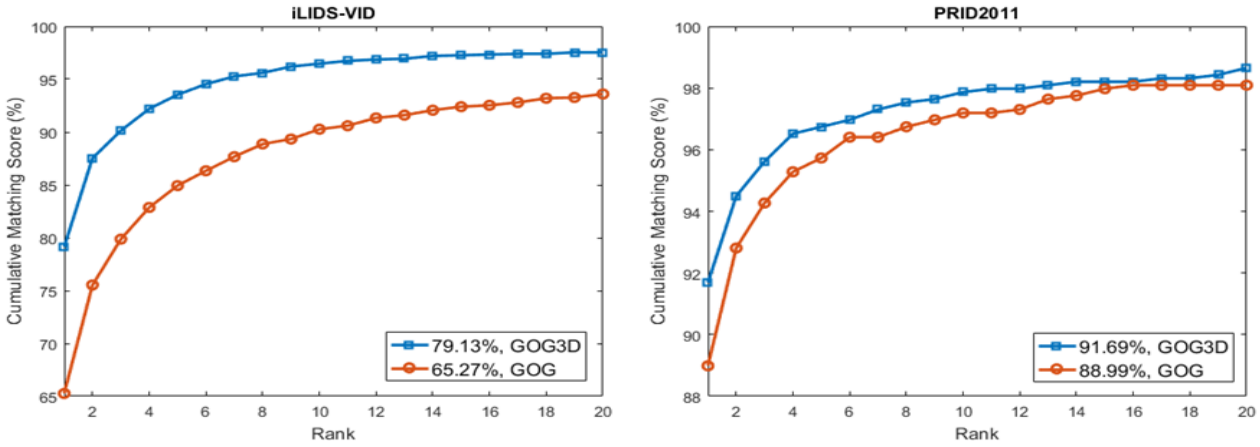


**FIGURE 4.** CMC curves using GOG and GOG3D features on iLIDS-VID and PRID2011 datasets.

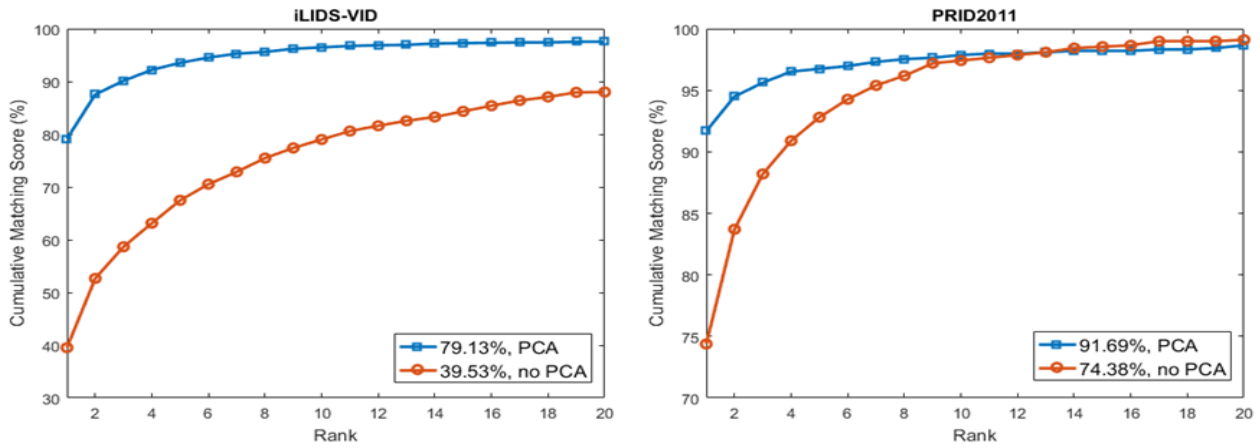


**FIGURE 5.** CMC curves of the results with and without performing PCA on iLIDS-VID and PRID2011 datasets.

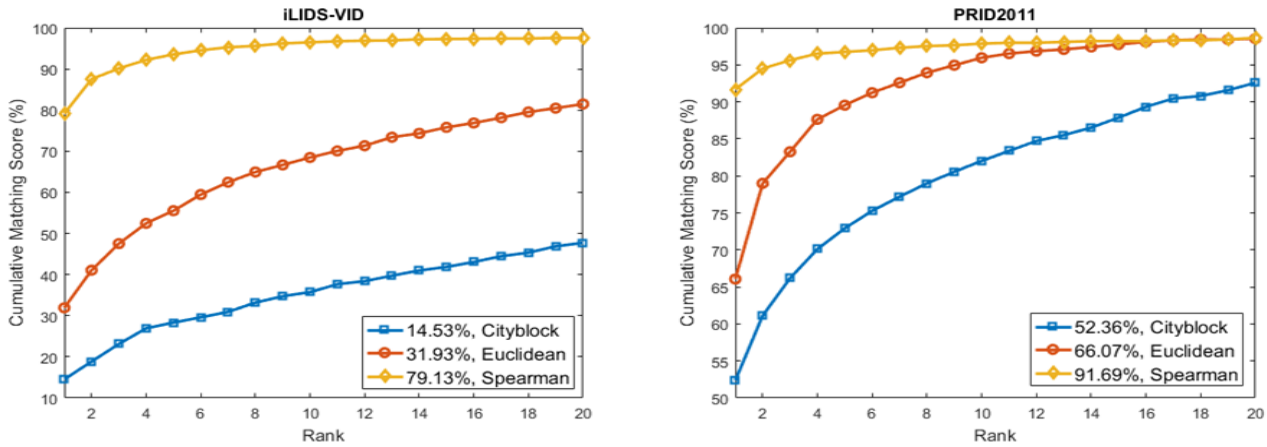


**FIGURE 6.** CMC curves using Cityblock, Euclidean and Spearman distances on iLIDS-VID and PRID2011 datasets.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2896779, IEEE Access



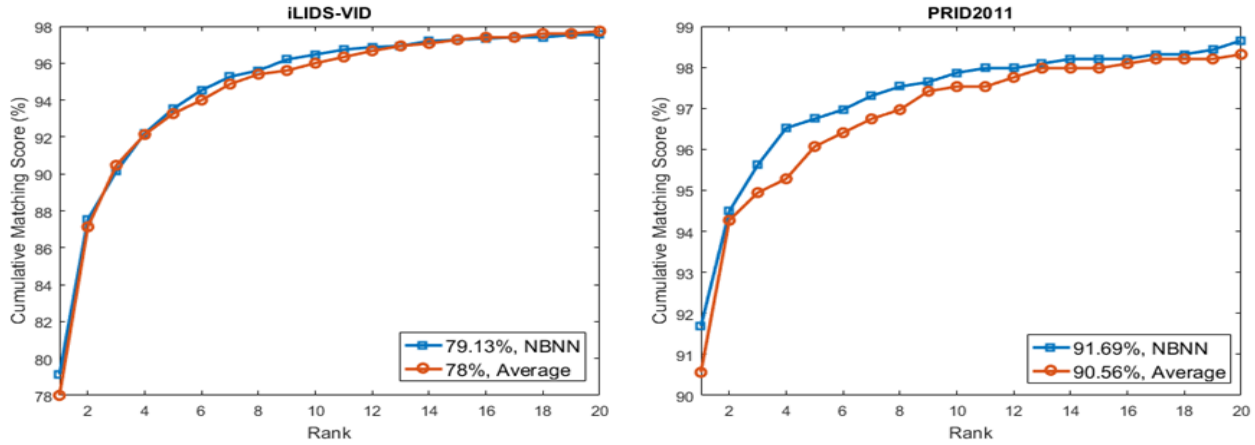


**FIGURE 7.** CMC curves comparing NBNN matching vs. feature average-pooling on iLIDS-VID and PRID2011 datasets.

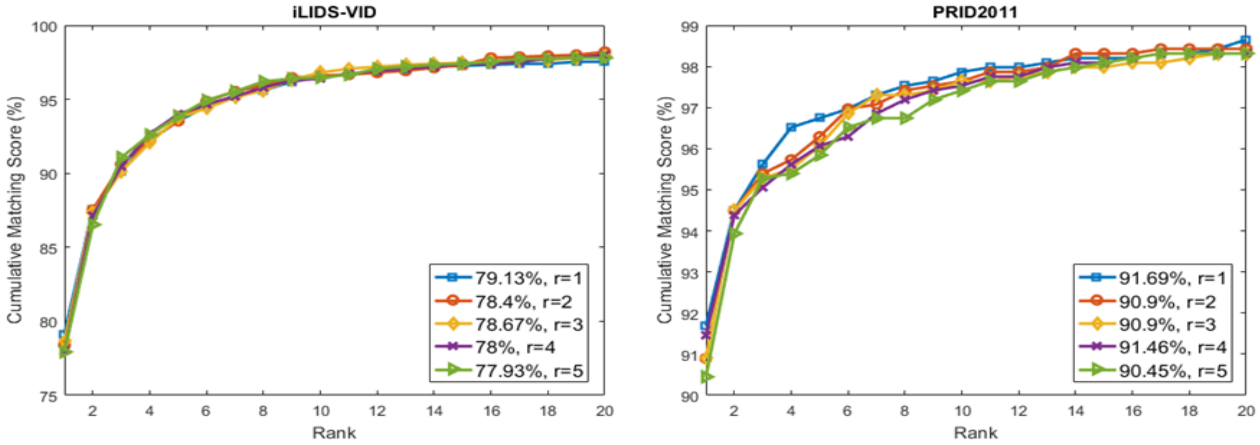


**FIGURE 8.** CMC curves obtained by varying the number of nearest neighbors r from 1 to 5 on iLIDS-VID and PRID2011 datasets.

ages the use of $r = 1$ to avoid unnecessary added complexity.

### *E. COMPUTATIONAL COMPLEXITY*

The proposed system was implemented in Matlab on a desktop PC with Intel Xeon CPU @ 3.60GHz and 64GB RAM. As learning is avoided altogether in this system, we report the matching time similarly to previous related works [24], [37]. For the testing phase, the time taken to compute the similarity between 2 tracklets is approximately 0.015 seconds. Therefore, the cost of finding the match for a given probe depends on the size of the gallery set. For instance, for PRID2011 dataset, it takes almost 1.35 seconds to generate a ranking list of the gallery elements with respect to a given probe. This indicates that the proposed method is in fact efficient.

## V. CONCLUSION

We have presented in this paper a novel learning-free method for unsupervised video-based person re-identification. By regarding video person re-id as a classification problem, we have adapted the NBNN classifier endowed with Spearman rank correlation coefficient as a similarity kernel into the set-based matching scenario. We have also extended state-of-the-art image descriptor GOG into 3 dimension achieving notable improvement in accuracy. An evaluation of the proposed method on 3 public benchmarks was conducted achieving outstanding results on 2 challenging small datasets, and competitive results on the large-scale benchmark. Different components of the proposed system were also thoroughly analyzed highlighting the amount each of them contributes to the performance improvement.

In future work, the proposed matching method could be embedded into a deep architecture to produce better feature representation that accounts for frames misalignment and scales better to large datasets.

## REFERENCES

[1] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in Computer Vision and Pattern Recognition (CVPR), 2012.

[2] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[3] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in European Conference on Computer Vision (ECCV), 2014.

[4] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[5] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in Computer Vision and Pattern Recognition (CVPR), 2018.

[6] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in European Conference on Computer Vision (ECCV), 2018.

[7] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in Computer Vision and Pattern Recognition (CVPR), 2018.

[8] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in Computer Vision and Pattern Recognition (CVPR), 2016.

[9] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in Computer Vision and Pattern Recognition (CVPR), 2018.

[10] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in Computer Vision and Pattern Recognition (CVPR), 2018.

[11] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in Computer Vision and Pattern Recognition (CVPR), 2018.

[12] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in International Conference on Computer Vision (ICCV), 2017.

[13] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in Computer Vision and Pattern Recognition (CVPR), 2017.

[14] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in Computer Vision and Pattern Recognition (CVPR), 2017.

[15] Y. Chen, X. Zhu, and S. Gong, "Deep association learning for unsupervised video person re-identification," in British Machine Vision Conference (BMVC), 2018.

[16] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI), 2018.

[17] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in International Conference on Computer Vision (ICCV), 2015.

[18] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in European Conference on Computer Vision (ECCV), 2016.

[19] C. Riachy and A. Bouridane, "Person re-identification: Attribute-based feature evaluation," in IEEE World Symposium on Applied Machine Intelligence and Informatics (SAMI), 2018.

[20] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in Computer Vision and Pattern Recognition (CVPR), 2008.

[21] S. McCann and D. G. Lowe, "Local naive bayes nearest neighbor for image classification," in Computer Vision and Pattern Recognition (CVPR), 2012.

[22] J. Weng, C. Weng, and J. Yuan, "Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition," in Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[23] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in Computer Vision and Pattern Recognition Workshops (CVPRW), 2012.

[24] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in IEEE International Conference on Computer Vision (ICCV), 2017.

[25] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in Computer Vision and Pattern Recognition (CVPR), 2017.

[26] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in Computer Vision and Pattern Recognition (CVPR), 2018.

[27] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in Computer Vision and Pattern Recognition (CVPR), 2016.

[28] W. Huang, C. Liang, Y. Yu, Z. Wang, W. Ruan, and R. Hu, "Video-based person re-identification via self paced weighting," in AAAI Conference on Artificial Intelligence, 2018.

[29] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in Computer Vision and Pattern Recognition (CVPR), 2018.

[30] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in Computer Vision and Pattern Recognition (CVPR), 2017.

[31] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in Computer Vision and Pattern Recognition (CVPR), 2016.

[32] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in Computer Vision and Pattern Recognition (CVPR), 2018.

[33] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in European Conference on Computer Vision (ECCV), 2014.

[34] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in International Conference on Computer Vision (ICCV), 2015.

[35] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person reidentification," in Computer Vision and Pattern Recognition (CVPR), 2017.

[36] S. Bak, P. Carr, and J.-F. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in European Conference on Computer Vision (ECCV), 2018.

[37] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, "Dynamic label graph matching for unsupervised video re-identification," in International Conference on Computer Vision (ICCV), 2017.

[38] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in European Conference on Computer Vision (ECCV), 2018.

[39] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong, "Person re-identification by unsupervised video matching," Pattern Recognition, vol. 65, pp. 197–210, 2017.

[40] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE transactions on pattern analysis and machine intelligence (TPAMI), vol. 32, no. 9, pp. 1627–1645, 2010.

[41] A. Dehghan, S. Modiri Assari, and M. Shah, "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in Computer Vision and Pattern Recognition (CVPR), 2015.

[42] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in Computer Vision and Pattern Recognition (CVPR), 2016.

[43] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in Conference on Computer Vision and Pattern Recognition (CVPR), 2013.

[44] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in Scandinavian Conference on Image Analysis (SCIA), 2011.

[45] F. M. Khan and F. Bremond, "Unsupervised data association for metric learning in the context of multi-shot person re-identification," in Advanced Video and Signal Based Surveillance (AVSS), 2016.

[46] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in European Conference on Computer Vision (ECCV), 2018.

[47] F. M. Khan and F. Brèmond, "Multi-shot person re-identification using part appearance mixture," in Winter Conference on Applications of Computer Vision (WACV), 2017.